

# Математические методы интеллектуального анализа данных

Неделько В. М.

Институт математики СО РАН, г. Новосибирск  
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».  
Лекция 1.

# Научное направление

Курс «Теория статистических решений» (д.т.н. В.Б. Бериков, к.ф.-м.н. В.М. Неделько) знакомит с научным направлением, известным как интеллектуальный анализ данных, а также машинное обучение, data mining, распознавание образов. Основной задачей данной научной области является разработка и обоснование методов автоматического поиска закономерностей в данных для получения новых знаний и прогнозирования.

Будут рассмотрены как классические методы дискриминантного, регрессионного и кластерного анализа, прогнозирования временных рядов, так и современные, в частности, нейронные сети, логистическая регрессия, композиции.

# Правила аттестации

Оценка за экзамен получается усреднением независимых оценок за первый и второй семестры.

По согласованию с деканатом возможен зачёт курса как полугодового.

Варианты аттестации:

- по билетам,
- статья на [machinelearning.ru](https://machinelearning.ru)
- демонстрация практического использования методов анализа данных, например, путём решения задач [kaggle.com](https://kaggle.com)

# Литература

- Г.С. Лбов. Анализ данных и знаний : учебное пособие. – Новосибирск : Изд-во НГТУ, 2001. – 86 с.
- В.М. Неделько. Основы статистических методов машинного обучения. Учебное пособие. НГТУ. 2010. 72 с.
- К.В. Воронцов. Машинное обучение (курс лекций) – 2009. <http://www.machinelearning.ru/>,  
<http://www.ccas.ru/voron/teaching.html>.
- А.Г. Дьяконов. Анализ данных, обучение по прецедентам, логические игры, системы weka, rapidminer и matlab. Учебное пособие.

## Литература дополнительная

- Лбов Г. С., Старцева Н. Г. Логические решающие функции и вопросы статистической устойчивости решений. 1999. 211 с.
- Лбов Г. С., Бериков В. Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации.
- Загоруйко Н. Г. Прикладные методы анализа данных и знаний, 1999. – 270 с.
- В.М. Неделько. Основы математической статистики: методы анализа данных. Учебно-методическое пособие. НГТУ. 2008. 44 с.
- В.М. Неделько. Основы теории вероятностей. Учебное пособие. НГТУ. 2011. 116 с.

# Ресурсы

- <http://www.machinelearning.ru/>
- <http://kaggle.com/>
- <http://archive.ics.uci.edu/ml/datasets.html>

## Задача Iris (Репозиторий UCI)

Задача состоит в том, чтобы построить правило, позволяющее по внешним признакам различать три вида (сорта) ирисов (цветы):

- 1 – Iris Setosa,
- 2 – Iris Versicolour,
- 3 – Iris Virginica.

В качестве измеряемых характеристик используются:

- $X_1$  – длина чашелистика (sepal length),
- $X_2$  – длина лепестка (petal length).

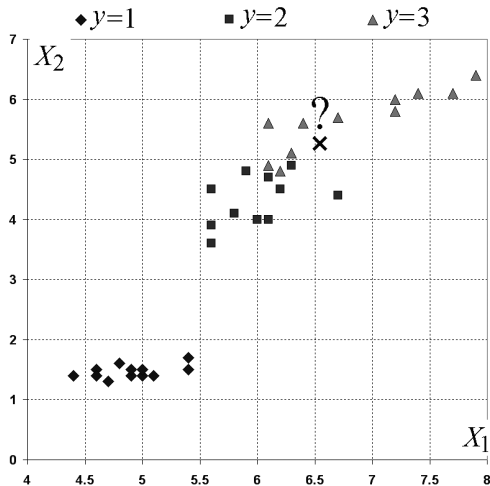
Необходимо построить функцию  $f$ , сопоставляющую любой паре  $(x_1, x_2)$  значений переменных  $X_1, X_2$  некоторое значение  $y$  переменной  $Y$ .

## Фрагмент таблицы данных

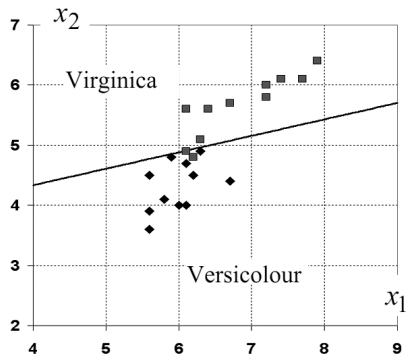
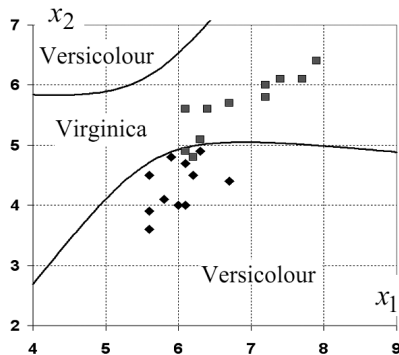
$i$	$x_1^i$	$x_2^i$	$y^i$	$i$	$x_1^i$	$x_2^i$	$y^i$
1	5,1	1,4	1	13	6	4	2
2	4,9	1,4	1	14	6,1	4,7	2
3	4,7	1,3	1	15	5,6	3,6	2
4	4,6	1,5	1	25	6,7	5,7	3
5	5	1,4	1	26	7,2	6	3
6	5,4	1,7	1	27	6,2	4,8	3
7	4,6	1,4	1	25	6,7	5,7	3



# Визуализация данных



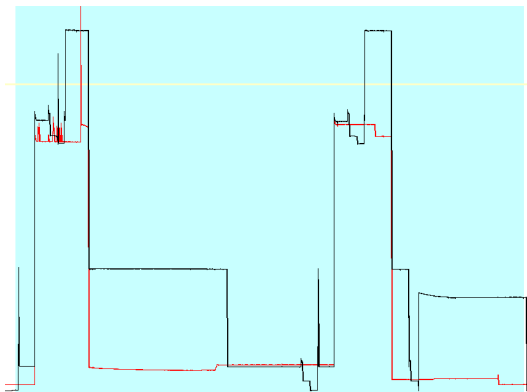
# Разделяющие кривые



## Примеры задач с kaggle.com

- Предсказание свойств белков.
- Оценивание школьных сочинений (эссе).
- Распознавание голосовых сигналов китов.
- Распознавание пользователя по показаниям акселерометра смартфона.
- Прогнозирование длительности полёта.
- Определение потребителя электроэнергии:

# Потребление электроэнергии по времени



# Задачи построения статистических решений

- Распознавание образов (классификация «с учителем»)
- Регрессионный анализ (восстановление зависимостей)
- Кластерный анализ (таксономия, автоматическая группировка, классификация «без учителя»)
- Задача упорядочивания объектов
- Задача обнаружения закономерностей
- Прогнозирование временных рядов
- Планирование эксперимента
- Поиск глобального экстремума
- Анализ клиентских сред
- Тематическое моделирование

## «Тэги»

- Искусственный интеллект.
- Нейронные сети.
- Экспертные системы.
- Распознавание образов.
- Интеллектуальный анализ данных.
- Коллаборативная фильтрация.
- Кредитный скоринг.

## Метод прецедентов

Для объекта, который нужно классифицировать, находятся наиболее похожие на него объекты, для которых целевой признак известен.

Прогноз осуществляется на основе «голосования» по прецедентам.

## Варианты метода прецедентов

- $K$ -ближайших соседей (прогноз на основе объектов, ближайших к исследуемому)
- Парзеновское окно (прогноз на основе объектов, попавших в окрестность исследуемого объекта)
- Ядровое сглаживание (вклад объекта в прогноз есть некоторая функция расстояния)

Сглаживание с использованием ядер есть наиболее общий вариант метода прецедентов.



## Свойства метода прецедентов

Метод прецедентов, несмотря на естественность и очевидность, на практике работает относительно плохо.

Недостатки метода:

- равное использование всех переменных, неустойчивость к «шумам», особенно при большой размерности,
- неиспользование возможной независимости переменных,
- неиспользование априорных знаний.

Существуют методы, существенно более эффективные (на большинстве практических задач).

## Качество решающей функции

- Насколько правильно решающая функция будет классифицировать «новые» объекты?
- Проблема «переобучения»: достаточно сложная решающая функция всегда точно классифицирует обучающую выборку, при этом «новые» объекты могут классифицироваться сколь угодно плохо.
- Модельный пример. Известно, что в урне белые и чёрные шары. Извлекли 10 шаров, все оказались белыми. Какой прогноз о цвете следующего шара?

# Границы применимости вероятностной постановки

Существуют также и невероятностные постановки задачи анализа данных (см. пособие по теории вероятностей).

Вероятностная постановка является достаточно общей и позволяет успешно решать подавляющее большинство практических задач. Однако требуется аккуратность.

Рассмотрим задачу, известную как «парадокс конвертов».

Игроку предлагается выбрать один из двух одинаковых на вид запечатанных конвертов с деньгами, причём известно, что сумма в одном из них в 10 раз больше, чем в другом.

Игроку разрешается вскрыть один конверт, после чего решить, забрать его или конверт, оставшийся запечатанным.

## «Парадокс конвертов»

Предположим, что игрок открыл наугад взятый конверт и обнаружил в нём 10 долларов.

Получается, что в другом конверте можно в равной степени ожидать сумму в 1 доллар или в 100 долларов. В среднем получаем 50,5.

Следует ли из этих рассуждений, что игроку выгоднее отказаться от первоначально выбранного конверта и забрать себе другой конверт?

Можно ли считать, что математическое ожидание выигрыша в этом случае (для рассмотренного примера) составит 50,5 долларов?